ddbjupdt@ddbj.nig.ac.jp
03/06/02 04:20 AM
Please respond to ddbjupdt

To: bmcleod@morganlewis.com
cc: ytateno@genes.nig.ac.jp, hsugawar@genes.nig.ac.jp, ddbjupdt@ddbj.nig.ac.jp
Subject: Re: public release date: AB008430

Dear Dr. Bonnie Weiss McLeod

The sequence data with accession number AB008430 were released from the DNA Data Bank of Japan (DDBJ) on Jan. 10 1998 in order to make them public.

DDBJ is in collaboration with the EMBL Nucleotide Sequence Databasein Europe and GenBank in USA to form and function as the International Nucleotide Sequence
Databases.

We take no responsibility for the priority and property issues for the submitted data. We simply inform you of the releasing date on request.

We appreciate your understanding and cooperation.

Sincerely yours,

Yoshio Tateno, Ph.D.
DNA Data Bank of Japan
National Institute of Genetics

>Date: Mon, 4 Mar 2002 17:32:50 +0900 (JST)
>From: ddbjupdt@ddbj.nig.ac.jp
>Subject: Re: public release date: AB008430
>To: bmcleod@morganlewis.com, pham@ncbi.nlm.nih.gov
>Cc: ddbjupdt@ddbj.nig.ac.jp

>Dear Sir,
>
>DNA Data Bank of Japan (DDBJ) has received your message at its update email
>address by way of GenBank.
>Your update message will be handled as soon as possible and
>in the order received. We appreciate your bringing this to our attention.
>
>Sincerely yours,
>DDBJ update
>
>>Date: Fri, 1 Mar 2002 11:46:14 -0500 (EST)
>>From: Vyvy Pham <pham@ncbi.nlm.nih.gov>
>>Subject: public release date: AB008430
>>To: ddbjupdt@ddbj.nig.ac.jp
>
>>
>>Dear DDBJ,
>>
>>I am forwarding a release date request for a patent inquiry.
>>Please reply directly to the user. Thank you for your help.
>>
>>Regards,
>>Vyvy Pham

>>GenBank User Services
>>
>>------------- Begin Forwarded Message -------------
>>
>>X-Server-Uuid: 76d36b76-3d48-11d4-a2af-00508bc764a5
>>Subject: public release date
>>To: info@ncbi.nlm.nih.gov
>>From: bmcleod@morganlewis.com
>>Date: Fri, 1 Mar 2002 09:24:48 -0500
>>X-MIMETrack: Serialize by Router on COLDGTW01/SVR/MLBLaw(Release 5.0.8 |June
>18,
>>2001) at 03/01/2002 09:24:50 AM
>>MIME-Version: 1.0
>>X-WSS-ID: 106150B8632552-01-01
>>Content-Transfer-Encoding: 7bit
>>X-Virus-Scanned: by amavisd-milter (http://amavis.org/)
>>
>>Hello,
>>I need to know the first date of public availability for the following
>>sequence record: Accession Number AB008430
>>The Journal section of the GenBank report says it was submitted on
>>22-10-97, but the Locus/last date of modification is listed as 13-02-99
>>Please let me know if you require further information to check the history
>>of the record.
>>Thanks
>>Bonnie Weiss McLeod
>>Morgan, Lewis & Bockius, LLP
>>202-739-6150 (phone)
>>202-739-3001 (fax)
>>bmcleod@morganlewis.com
>>
>>
>>**
>>This e-mail message is intended only for the personal use of the
recipient(s)
>>named above. This message may be an attorney-client communication and as
such
>>privileged and confidential. If you are not an intended recipient, you may
not
>>review, copy or distribute this
>>message. If you have received this communication in error, please notify us
>>immediately by e-mail and delete the original message.
>>**coldsc
n0
1

"Vyvy Pham" <pham@ncbi.nlm.nih.gov>

03/01/02 11:47 AM
Please respond to "Vyvy Pham"

To: bmcleod@morganlewis.com
cc: pham@ncbi.nlm.nih.gov
Subject: public release date

Dear Colleague,

We received your inquiry about the release date of AB008430
at the National Center for Biotechnology Information.

GenBank collaborates with the EMBL database in Europe and the DDBJ database in Japan to produce and distribute the International DNA Sequence Databases. A sequence is submitted to only one of the three databases, then shared among the three after the record is released. Information about the history of a record, including the first date of public release, can only be provided by the database to which the original submission was made.

Since the accession number about which you have inquired was originally submitted to DDBJ, we have forwarded your inquiry to them. They will reply directly to you regarding the release date. If you have further questions about that accession, please contact DDBJ at ddbjupdt@ddbj.nig.ac.jp .

The International DNA Sequence Databases take no responsibility for making any determination of the priority issues for patent claims. We are able only to inform you, to the best of our knowledge, of the release date of the pertinent sequence into the public database. We appreciate your understanding and cooperation.

Regards,

Vyvy Pham
GenBank User Services

---------- Begin Forwarded Message ----------

X-Server-Uuid: 76d36b76-3d48-11d4-a2af-00508bc764a5
Subject: public release date
To: info@ncbi.nlm.nih.gov
From: bmcleod@morganlewis.com
Date: Fri, 1 Mar 2002 09:24:48 -0500
X-MIMETrack: Serialize by Router on COLDGTW01/SVR/MLBLaw(Release 5.0.8 |June 18,
2001) at 03/01/2002 09:24:50 AM
MIME-Version: 1.0
X-WSS-ID: 106150B8632552-01-01
Content-Transfer-Encoding: 7bit
X-Virus-Scanned: by amavisd-milter (http://amavis.org/)

Hello,
I need to know the first date of public availability for the following sequence record: Accession Number AB008430
The Journal section of the GenBank report says it was submitted on

22-10-97, but the Locus/last date of modification is listed as 13-02-99
Please let me know if you require further information to check the history
of the record.
Thanks
Bonnie Weiss McLeod
Morgan, Lewis & Bockius, LLP
202-739-6150 (phone)
202-739-3001 (fax)
bmcleod@morganlewis.com

---------- End Forwarded Message ----------

1: Biochem Biophys Res Commun 1997 Dec 18;241(2):369-75

Molecular cloning and characterization of CDEP, a novel human protein containing the ezrin-like domain of the band 4.1 superfamily and the Dbl homology domain of Rho guanine nucleotide exchange factors.

Koyano Y, Kawamoto T, Shen M, Yan W, Noshiro M, Fujii K, Kato Y.

Department of Biochemistry, Hiroshima University School of Dentistry, Japan.

A cDNA for a novel human protein named CDEP was cloned using the subtractive hybridization method between dedifferentiated cartilage cells and overtly differentiated cartilage cells. CDEP cDNA contained an open reading frame encoding 1,045 amino acids in a total length of 3.4 kb. The deduced amino acid sequence revealed that a single polypeptide contained the ezrin-like domain, which is found in cytoskeleton-associated proteins of the band 4.1 superfamily, and the Dbl homology (DH) and pleckstrin homology (PH) domains, which are conserved in the Rho GEF (guanine nucleotide exchange factor) family. Northern blot analysis demonstrated that CDEP mRNA was expressed not only in the differentiated chondrocytes but also in various fetal and adult tissues. Since members of the band 4.1 superfamily and the Rho GEF family are crucial for microfilament organization, the novel protein CDEP may be involved in the adhesion, proliferation, and differentiation of some cell types including chondrocytes via changes in the cytoskeleton.

PMID: 9425278 [PubMed - indexed for MEDLINE]

NCBI **GenBank Overview**

PubMed Entrez BLAST OMIM Books Taxonomy Structure

What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research* 2000 Jan 1;28(1):15-8). There are approximately 15,850,000,000 bases in 14,976,000 sequence records as of December 2001 (see GenBank growth statistics). As an example, you may view the record for a *Saccharomyces cerevisiae* gene. The complete release notes for the current version of GenBank are available. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which is comprised of the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

Submissions to GenBank

Many journals require submission of sequence information to a database prior to publication so that an accession number may appear in the paper. NCBI has a WWW form, called BankIt, for convenient and quick submission of sequence data. Sequin, NCBI's stand-alone submission software for MAC, PC, and UNIX platforms, is also available by FTP. When using Sequin, the output files for direct submission should be sent to GenBank by electronic mail.

There are specialized, streamlined procedures for batch submissions of sequences, such as EST, STS, and HTG sequences.

Updating or Revising a Sequence

Revisions or updates to GenBank entries can be made at any time and can be accepted as BankIt or Sequin files or as the text of an e-mail message. Be sure to give the accession number of the sequence to be updated in the subject line. Send it to:

update@ncbi.nlm.nih.gov

Access to GenBank

GenBank is available for searching at NCBI via several

methods.

The GenBank database is designed to provide and encourage access within the scientific community to the most up to date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

▸ New Developments

NCBI is continuously developing new tools and enhancing existing ones to improve both submission and access to GenBank. The easiest way to keep abreast of these and other developments is to check the "What's New" section of the NCBI Web page and to read the *NCBI News*, which is also available by free subscription.

Revised January 7, 2002

NCBI **Submit to GenBank**

Submitting Sequence Data to GenBank

The most important source of new data for GenBank® is direct submissions from scientists. GenBank depends on its contributors to help keep the database as comprehensive, current, and accurate as possible. NCBI provides timely and accurate processing and biological review of new entries and updates to existing entries, and is ready to assist authors who have new data to submit.

NOTE: The 'Authorin' submission tool and the E-mail submission form were phased out on December 31, 1998, and submissions made with those tools are no longer accepted as of that date. Instead, please use the improved submission tools, BankIt and Sequin, described below.

Submit now!!

Sequin
Stand-alone sequence submission tool

BankIt
For quick and simple submissions

VecScreen
Vector contamination screening tool

GenBank

GenBank
overview of the database

Search GenBank
explore the data

Receiving an accession number for your manuscript

Most journals now expect that DNA and amino acid sequences that appear in articles will be submitted to a sequence database before publication. Soon after submission, you will receive an accession number from the database which you will be able to use in your article to refer to the sequence. Please be aware that it is only necessary to submit the sequence to one database, whichever one is most convenient, without regard for where the sequence may be published. Data exchange between GenBank, EMBL and DDBJ occurs daily. Sequence data submitted in advance of publication can be kept confidential if requested.

Below are described various ways of submitting DNA sequences to GenBank. Essentially, there are two principal ways, BankIt and Sequin. BankIt is a Web submission tool and recommended for simple submissions. With BankIt you can indicate coding regions on an mRNA along with a product and gene name. For more control over annotating your entry, segmented records, or very long entries, Sequin, a stand-alone

submission tool, is suggested.

GenBank will provide you with an accession number to identify your sequence, usually within two working days, if the submission is received via electronic mail. This accession number serves as confirmation that you have submitted your data, and allows the community to retrieve the data upon reading the journal article.

The accession number should be included in your manuscript, preferably in a footnote on the first page of the article, or as required by individual journal procedures.

BankIt - submitting via the WWW

NCBI has developed a WWW form, called BankIt, for convenient and quick submission of sequence data.

BankIt allows you to enter sequence information into a form, edit as necessary, and add biological annotation (e.g., coding regions, mRNA features). BankIt transforms your data into GenBank format for your review and when your record is completed, it can be submitted directly to GenBank. You have the option of adding information by using text boxes to describe in your own words the source of the sequence and its biological features. The GenBank annotation staff reviews the submitted textual information, incorporates it into the appropriate structured fields, and returns the record by e-mail for your review.

BankIt is compatible with Netscape clients for Unix, Macs, and PCs. In addition, Internet Explorer for the PC and Mac have successfully been used.

Sequin - stand-alone software for the Mac, PC/Windows, and UNIX

If you do not have access to the WWW, NCBI introduces a stand-alone submission program called Sequin.

Sequin is an interactive, graphically-oriented program based on screen forms and controlled vocabularies that guides you through the process of entering your sequence and providing biological and bibliographic annotation. Sequin is designed to simplify the sequence submission process and to provide graphical viewing and editing options. It incorporates robust error checking and accommodates very long sequences and complex annotations.

Special submissions - genomes, batch sequences, alignments

Sequin can be used for the submission of individual or small numbers of sequences. However, it was also designed to facilitate special types of submissions, and should be used instead of BankIt for the following types of submissions: genomes and other very long sequences; multiple sequences such as batch submissions and segmented sets; and

population/phylogenetic/mutation studies.

When preparing the submission of a genome, you can import the complete genome sequence into Sequin as well as a file containing the amino acid translations in FASTA format, if available. Sequin will automatically annotate the coding regions intervals based on the translations, and you can use Sequin to make further complex annotations. Sequin can also accept feature annotations in tab-delineated tables. Since the final submission file (*.sqn) will be quite large, please send it to the GenBank staff via FTP rather than by e-mail. To request a temporary FTP directory, please contact genomes@ncbi.nlm.nih.gov.

When preparing a submission that contains multiple sequences, you can import a single file containing all the sequences in FASTA format, or as alignments in FASTA+GAP, PHYLIP, or NEXUS format. In addition, for population/phylogenetic/mutation studies, you can annotate one sequence and propagate the features onto the other sequences. When you complete the submission and select the 'prepare submission' option in the 'File' menu, Sequin will prepare a single *.sqn file that contains all the sequences. Send the *.sqn file by e-mail to:

gb-sub@ncbi.nlm.nih.gov .

If you are submitting two or more Sequin files, each of which contains multiple sequences, send each *.sqn file in a separate e-mail message.

Please refer to the Sequin Quick Guide and documentation for additional information, both of which are accessible from the Sequin Web page.

Sending the Data to GenBank

When using BankIt, the prepared sequence entries are submitted directly to GenBank through the WWW.

When using Sequin, the output files for direct submission should be sent to GenBank by electronic mail to:

gb-sub@ncbi.nlm.nih.gov

As an alternative, the submission file can be copied to floppy disk and mailed to GenBank Submissions at:

GenBank Submissions
National Center for Biotechnology Information
National Library of Medicine
Bldg. 38A, Room 8N-803
Bethesda, MD 20894

Please label the disk with your name and file name and

indicate whether it is a PC or MAC disk.

Updates

NCBI processes update requests as well as new submissions. You can provide additional annotation, correct errors or omissions, or request the release of your "hold-until-published" record. BankIt or Sequin may be used for updates, or you can request changes as text in the body of an e-mail message. Be sure to give the accession number of the sequence to be updated along with all update information. Send it to:

update@ncbi.nlm.nih.gov

Submitters of a record maintain editorial control of that record. Any third party update information will be forwarded to the submitters of the record for review. Changes will be made to the record only at the submitters' request. If submitters can no longer be contacted, GenBank reserves the right to edit an entry to agree with the information presented in the original publication(s) cited in the entry.

Submission of ESTs, STSs and GSSs

Batches of ESTs (expressed sequence tags), STSs (sequence tagged sites), and GSSs (genome survey sequences) can be submitted via special streamlined procedures.

Submission of HTGS Records

The NCBI has developed a protocol for high throughput genome sequencing centers to use when they submit large genomic records (usually Cosmids or BACs). Specialized tools, including fa2htgs and a "genome center version" of Sequin, have been created to help such centers produce these submission files in a convenient way. The HTG page not only provides detailed submission instructions to genome centers, but also informs GenBank users how to access the HTG sequences.

Confidentiality

Some authors are concerned that the appearance of their data in GenBank prior to publication will compromise their work. GenBank will, upon request, withhold release of new submissions for a specified period of time. However, if a paper citing the sequence or accession number is published prior to the specified date, your sequence will be released upon publication.

In order to prevent the delay in the appearance of published sequence data, we urge authors to inform us of the appearance of the published data. As soon as it is available, please send the full publication data--all authors, title, journal, volume, pages and

date--to the following address:

update@ncbi.nlm.nih.gov

Submission of SNPs and other polymorphism data

Data on genetic variation in humans and other organisms can be submitted to the NCBI Database of Single Nucleotide Polymorphisms (dbSNP). Entries include single nucleotide polymorphisms (SNPs), small-scale insertion/deletions, polymorphic repetitive elements, and microsatellite variation. dbSNP is a separate resource from the GenBank database, and submissions do not receive GenBank accessions as noted above. However, dbSNP entries do receive dbSNP identifiers and contain links to associated GenBank records. Further information about submitting data is accessible from the sidebar of the dbSNP home page.

Revised January 7, 2002

NCBI **Sample GenBank Record** SITE MAP

PubMed Entrez BLAST OMIM Taxonomy Structure

GenBank Flat File Format

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      baker's yeast.
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales;
            Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  MEDLINE   95176709
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  MEDLINE   96194260
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"
                     /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
```

TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE
TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV
YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG
DVIYFNFEVVSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTNVSLEFTNSSQ
DHDWVKFQSSNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFNIIGMDSKITHSNHSA
NATSTRSSHHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA
CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKPNQENATPLN
NPFDDDASSYDDTSIARRLAALNTLKLDNHSATESDISSVDEKRDSLSGMNTYNDQFQ
SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTGNLSPVSDIVRDS
YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK
HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
VDFSNKSNVNVGQVKDIHGRIPEML"

gene complement(3300..4037)
/gene="REV7"

CDS complement(3300..4037)
/gene="REV7"
/codon_start=1
/product="Rev7p"
/protein_id="AAA98667.1"
/db_xref="GI:1293616"
/translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
LISGDDKILNGVYSQYEEGESIFGSLF"

BASE COUNT 1510 a 1074 c 835 g 1609 t

ORIGIN

```
      1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
     61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
    121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
    181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
    241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
    301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
    361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
    421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
    481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
    541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
    601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta
    661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag
    721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa
    781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt tccaatgata
    841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc gacttaccga
    901 gctggctttc gtttgactct agttctagaa cgttctcagg tgaaccttct tctgacttac
    961 tatctgatgc gaacaccacg ttgtatttca atgtaatact cgagggtacg gactctgccg
   1021 acagcacgtc tttgaacaat acataccaat ttgttgttac aaaccgtcca tccatctcgc
   1081 tatcgtcaga tttcaatcta ttggcgttgt taaaaaacta tggttatact aacggcaaaa
   1141 acgctctgaa actagatcct aatgaagtct tcaacgtgac ttttgaccgt tcaatgttca
   1201 ctaacgaaga atccattgtg tcgtattacg gacgttctca gttgtataat gcgccgttac
   1261 ccaattggct gttcttcgat tctggcgagt tgaagtttac tgggacggca ccggtgataa
   1321 actcggcgat tgctccagaa acaagctaca gttttgtcat catcgctaca gacattgaag
   1381 gattttctgc cgttgaggta gaattcgaat tagtcatcgg ggctcaccag ttaactacct
   1441 ctattcaaaa tagtttgata atcaacgtta ctgacacagg taacgtttca tatgacttac
   1501 ctctaaacta tgtttatctc gatgacgatc ctatttcttc tgataaattg ggttctataa
   1561 acttattgga tgctccagac tgggtggcat tagataatgc taccatttcc gggtctgtcc
   1621 cagatgaatt actcggtaag aactccaatc ctgccaattt ttctgtgtcc atttatgata
   1681 cttatggtga tgtgatttat ttcaacttcg aagttgtctc cacaacggat ttgtttgcca
   1741 ttagttctct tcccaatatt aacgctacaa ggggtgaatg gttctcctac tatttttgc
   1801 cttctcagtt tacagactac gtgaatacaa acgtttcatt agagtttact aattcaagcc
   1861 aagaccatga ctgggtgaaa ttccaatcat ctaatttaac attagctgga gaagtgccca
```

```
1921 agaatttcga caagctttca ttaggtttga aagcgaacca aggttcacaa tctcaagagc
1981 tatattttaa catcattggc atggattcaa agataactca ctcaaaccac agtgcgaatg
2041 caacgtccac aagaagttct caccactcca cctcaacaag ttcttacaca tcttctactt
2101 acactgcaaa aatttcttct acctccgctg ctgctacttc ttctgctcca gcagcgctgc
2161 cagcagccaa taaaacttca tctcacaata aaaaagcagt agcaattgcg tgcggtgttg
2221 ctatcccatt aggcgttatc ctagtagctc tcatttgctt cctaatattc tggagacgca
2281 gaagggaaaa tccagacgat gaaaacttac cgcatgctat tagtggacct gatttgaata
2341 atcctgcaaa taaaccaaat caagaaaacg ctacaccttt gaacaacccc tttgatgatg
2401 atgcttcctc gtacgatgat acttcaatag caagaagatt ggctgctttg aacactttga
2461 aattggataa ccactctgcc actgaatctg atatttccag cgtggatgaa aagagagatt
2521 ctctatcagg tatgaataca tacaatgatc agttccaatc ccaaagtaaa gaagaattat
2581 tagcaaaacc cccagtacag cctccagaga gcccgttctt tgacccacag aataggtctt
2641 cttctgtgta tatggatagt gaaccagcag taaataaatc ctggcgatat actggcaacc
2701 tgtcaccagt ctctgatatt gtcagagaca gttacggatc acaaaaaact gttgatacag
2761 aaaaactttt cgatttagaa gcaccagaga aggaaaaacg tacgtcaagg gatgtcacta
2821 tgtcttcact ggacccttgg aacagcaata ttagcccttc tcccgtaaga aaatcagtaa
2881 caccatcacc atataacgta acgaagcatc gtaaccgcca cttacaaaat attcaagact
2941 ctcaaagcgg taaaaacgga atcactccca caacaatgtc aacttcatct tctgacgatt
3001 ttgttccggt taaagatggt gaaaattttt gctgggtcca tagcatggaa ccagacagaa
3061 gaccaagtaa gaaaaggtta gtagattttt caaataagag taatgtcaat gttggtcaag
3121 ttaaggacat tcacggacgc atcccagaaa tgctgtgatt atacgcaacg atattttgct
3181 taattttatt ttcctgtttt atttttttatt agtggtttac agatacccta tattttattt
3241 agtttttata cttagagaca tttaatttta attccattct tcaaatttca tttttgcact
3301 taaaacaaag atccaaaaat gctctcgccc tcttcatatt gagaatacac tccattcaaa
3361 attttgtcgt caccgctgat taatttttca ctaaactgat gaataatcaa aggccccacg
3421 tcagaaccga ctaaagaagt gagttttatt ttaggaggtt gaaaaccatt attgtctggt
3481 aaatttttcat cttcttgaca tttaacccag tttgaatccc tttcaatttc tgctttttcc
3541 tccaaactat cgaccctcct gtttctgtcc aacttatgtc ctagttccaa ttcgatcgca
3601 ttaataactg cttcaaatgt tattgtgtca tcgttgactt taggtaattt ctccaaatgc
3661 ataatcaaac tatttaagga agatcggaat tcgtcgaaca cttcagtttc cgtaatgatc
3721 tgatcgtctt tatccacatg ttgtaattca ctaaaatcta aaacgtattt ttcaatgcat
3781 aaatcgttct ttttattaat aatgcagatg gaaaatctgt aaacgtgcgt taatttagaa
3841 agaacatcca gtataagttc ttctatatag tcaattaaag caggatgcct attaatggga
3901 acgaactgcg gcaagttgaa tgactggtaa gtagtgtagt cgaatgactg aggtgggtat
3961 acatttctat aaaataaaat caaattaatg tagcatttta agtataccct cagccacttc
4021 tctacccatc tattcataaa gctgacgcaa cgattactat ttttttttc ttcttggatc
4081 tcagtcgtcg caaaaacgta taccttcttt ttccgacctt ttttttagct ttctggaaaa
4141 gtttatatta gttaaacagg gtctagtctt agtgtgaaag ctagtggttt cgattgactg
4201 atattaagaa agtggaaatt aaattagtag tgtagacgta tatgcatatg tatttctcgc
4261 ctgtttatgt ttctacgtac ttttgattta tagcaagggg aaaagaaata catactattt
4321 tttggtaaag gtgaaagcat aatgtaaaag ctagaataaa atggacgaaa taaagagagg
4381 cttagttcat ctttttttcca aaaagcaccc aatgataata actaaaatga aaaggatttg
4441 ccatctgtca gcaacatcag ttgtgtgagc aataataaaa tcatcacctc cgttgccttt
4501 agcgcgtttg tcgtttgtat cttccgtaat tttagtctta tcaatgggaa tcataaattt
4561 tccaatgaat tagcaatttc gtccaattct ttttgagctt cttcatattt gctttggaat
4621 tcttcgcact tcttttccca ttcatctctt tcttcttcca aagcaacgat ccttctaccc
4681 atttgctcag agttcaaatc ggcctctttc agtttatcca ttgcttcctt cagtttggct
4741 tcactgtctt ctagctgttg ttctagatcc tggtttttct tggtgtagtt ctcattatta
4801 gatctcaagt tattggagtc ttcagccaat tgctttgtat cagacaattg actctctaac
4861 ttctccactt cactgtcgag ttgctcgttt ttagcggaca aagatttaat ctcgttttct
4921 ttttcagtgt tagattgctc taattctttg agctgttctc tcagctcctc atatttttct
4981 tgccatgact cagattctaa ttttaagcta ttcaatttct ctttgatc
//
```

Other Formats: FASTA | ASN.1 | Back to Top

Examples of other records that show a range of biological features

FIELD	COMMENTS

LOCUS

- **Locus Name**

The locus name was originally designed to help group entries with similar sequences: the first three characters usually designated the organism; the fourth and fifth characters were used to show other group designations, such as gene product; for segmented entries the last character was one of a series of sequential integers. (See GenBank release notes section 3.4.4 for more info.)

However, the ten characters in the locus name are no longer sufficient to represent the amount of information originally intended to be contained in the Locus name. The only rule now applied in assigning a Locus name is that it must be unique. For example, for GenBank records that have 6-character accessions (e.g., U12345), the locus name is usually the first letter of the genus and species name followed by the accession number. For 8-character character accessions (e.g., AF123456), the locus name is just the accession number.

The RefSeq database of reference sequences assigns formal locus names to each record, based on gene symbol. RefSeq is separate from the GenBank database, but contains cross-references to corresponding GenBank records.

Entrez Search Field: Accession Number [ACCN]
Search Tip: It is better to search for the actual accession number rather than the locus name, since the accessions are stable and locus names can change.

- **Sequence Length**

Number of nucleotide base pairs (or amino acid residues) in the sequence record.

There is no maximum limit on the size of a sequence that can be submitted to GenBank - you can submit a whole genome if you have a contiguous piece of sequence from a single molecule type. However, there is a limit of 350 kb on an individual GenBank record (with some exceptions, as noted in section 1.3.2 of the release notes for GenBank 112.0). That limit was agreed upon by the international collaboratoring sequence databases to facilitate handling of sequence data by various software programs. (For more information, see NCBI News articles on Complete Genomes and GenBank Enters Megabase Era.) The minimum length required for submission is 50 bp, although there might be some shorter records from past years.

Entrez Search Field:

Sequence Length [SLEN]
Search Tips: (1) To retrieve records within a range of lengths, use the colon as the range operator, e.g., 2500:2600[slen]. (2) To retrieve all sequences shorter than a certain number, use 2 as the lower bound, e.g., 2:100[slen]. (3) To retrieve all sequences longer than a certain number, use a series of 9's as the upper bound, e.g., 325000:99999999[slen].

- **Molecule Type**

The type of molecule that was sequenced.

Each GenBank record must contain contiguous sequence data from a single molecule type. The various molecule types are described in the Sequin documentation, and can include genomic DNA, genomic RNA, precursor RNA, mRNA (cDNA), ribosomal RNA, transfer RNA, small nuclear RNA, and small cytoplasmic RNA.

Entrez Search Field:

Properties [PROP]
Search Tip: Search term should be in the format: **biomol_genomic**, **biomol_mRNA**, etc. For more examples, view the Properties field in "Index" mode.

- **GenBank Division**

The GenBank database is divided into 17 divisions:

1. PRI - primate sequences
2. ROD - rodent sequences
3. MAM - other mammalian sequences
4. VRT - other vertebrate sequences
5. INV - invertebrate sequences
6. PLN - plant, fungal, and algal sequences
7. BCT - bacterial sequences
8. VRL - viral sequences
9. PHG - bacteriophage sequences
10. SYN - synthetic sequences
11. UNA - unannotated sequences
12. EST - EST sequences (expressed sequence tags)
13. PAT - patent sequences
14. STS - STS sequences (sequence tagged sites)
15. GSS - GSS sequences (genome survey sequences)
16. HTG - HTGS sequences (high throughput genomic sequences)
17. HTC - unfinished high-throughput cDNA sequencing

Some of the divisions contain sequences from specific groups of organisms, while others (EST, GSS, HTG, etc.) contain data generated by specific sequencing technologies from many different organisms. The **organismal divisions are historical and do not reflect the current NCBI Taxonomy**. Instead, they merely serve as a convenient way to divide GenBank into smaller pieces for those who want to FTP the database. Because of this, and because sequences from a particular organism can exist in technology-based divisions such as EST, HTG, etc., **the NCBI Taxonomy Browser should be used for retrieving all sequences from a particular organism**.

The divisions are also listed in section 3.3 of the GenBank release notes.

The RNA division of GenBank was removed in release 113.0 (August 1999). Sequences that were previously in the RNA division have been moved to the appropriate organismal division. (See section 1.3.2 of the GenBank 113.0 release notes for additional information.)

The HTC division was added to GenBank in release 123.0 (April 2001), and is described in Section 1.3.3 of the GenBank 123.0 release notes.

An 18th division, called CON, was added in release 115.0 (December 1999) but is not listed above because it is still experimental. Records in that division contain no sequence data. Instead, they contain instructions on

how to construct contigs from multiple GenBank records. See the Fall 1999 NCBI News and section 1.3.3 of GenBank 115.0 release notes for details.

Entrez Search Field:

Properties [PROP]
Search Tip: Search term should be in the format: *gbdiv_pri*, *gbdiv_est*, etc. For more examples, view the Properties field in "Index" mode. For example, to eliminate all sequences from a particular division, such as all ESTs, you can use a Boolean query formatted such as:

human[orgn] NOT gbdiv_est[prop]

For the reasons noted above, **do not use GenBank divisions to retrieve all sequences from a specific organism. Instead, use the NCBI Taxonomy Browser**.

- **Modification Date** The date in the LOCUS field is the **date of last modification**. In some cases, it might correspond to the release date, but there is no way to tell just by looking at the record. If you need to know the first date of public availability for a specific sequence record, send a message to info@ncbi.nlm.nih.gov. We will check the history of the record for you, and let you know the date of first public release. If the sequence was originally submitted to our collaborators at DDBJ or EMBL, rather than to GenBank, we will ask them to send the release date information to you. (See also notes re: date in the Direct Submission reference.)

  Entrez Search Field:

  Modification Date [MDAT]
  Search Tips: (1) Enter search term in the format: yyyy/mm/dd, e.g., 1999/07/25. (2) To retrieve records modified between two dates, use the colon as a range operator, e.g., 1999/07/25:1999/07/31[mdat]. (3) You can use the Publication Date [PDAT] field of Entrez to limit search results by the date on which records were added to the Entrez system. Publication date can be ranged just like the Modification Date.

DEFINITION

Brief description of sequence; includes information such as source organism, gene name/protein name, or some description of the sequence's function (if the sequence is non-coding). If the sequence has a coding region (CDS), description may be followed by a completeness qualifier, such as "complete cds." (See GenBank release notes section 3.4.5 for more info.)

Entrez Search Field: Title Word [TITL]
Search Tip: Although nucleotide definition lines follow a structured format, GenBank does not use a controlled vocabulary and authors determine the content of their records. Therefore, if a search for a specific term does not retrieve the desired records, try other terms that authors might have used, such synonyms, full spellings, or abbreviations. The 'related records' (or 'neighbors') function of Entrez also allows you to broaden your search by retrieving records with similar sequences, regardless of the descriptive terms used by the submitters.

ACCESSION

The unique identifier for a sequence record. An accession number applies to the complete record and is usually a combination of a letter(s) and numbers, such as a single letter followed by five digits (e.g., U12345), or two letters followed by six digits (e.g., AF123456). Accession numbers do not change, even if information in the record is changed at the author's request. Sometimes, however, an original accession number might become secondary to a newer accession number, if the authors make a new submission that combines previous sequences, or if for some reason a new submission supercedes an earlier record.

Records from the RefSeq database of reference sequences have a different accession number format that begins with two letters followed by an underscore bar and six digits, for example:

NT_123456 constructed genomic contigs
NM_123456 mRNAs
NP_123456 proteins
NC_123456 chromosomes

Note: compare accession number with Sequence Identifiers such as Version and GI for nucleotide sequences, and ProteinID and GI for amino acid sequences.

Entrez Search Field: Accession [ACCN]

Search Tip: The letters in the accession number can be written in upper or lower case. RefSeq accessions must contain an underscore bar between the letters and the numbers, e.g, NM_002111.

VERSION

A nucleotide sequence identification number that represents a single, specific sequence in the GenBank database. This identification number uses the accession.version format implemented by GenBank/EMBL/DDBJ in February 1999.

If there is any change to the sequence data (even a single base) the version number will be increased, e.g., U12345.1 --> U12345.2, but the accession portion will remain stable.

The accession.version system of sequence identifiers runs parallel to the GI number system. That is, when any change is made to a sequence, it receives a new GI number AND an increase to its version number.

For more information, see section 1.3.2 of the GenBank 111.0 release notes, and section 3.4.7 of the current GenBank release notes.

A Sequence Revision History tool is available to track the various gi numbers, version numbers, and update dates for sequences that appeared in a specific GenBank record (more information and example).

Entrez Search Field: Can use either Accession [ACCN] or UID

- **GI**

"GenInfo Identifier" sequence identification number, in this case, for the nucleotide sequence. If a sequence changes in any way, a new GI number will be assigned.

A separate GI number is also assigned to each protein translation within a nucleotide sequence record, and a new GI is assigned if the protein translation changes in any way (see below).

GI sequence identifiers run parallel to the new **accession.version** system of sequence identifiers. For more information, see the description of Version, above, and section 3.4.7 of the current GenBank release notes.

Entrez Search Field: UID

KEYWORDS

Word or phrase describing the sequence. If no keywords are included in the entry, the field contains only a period.

The Keyword field is present in sequence records primarily for historical reasons, and is not based on a controlled vocabulary. Keywords are generally present in older records. They are **not** included in newer records unless (1) they are not redundant with any feature, qualifier, or other information present in the record, or (2) the submitter specifically asks for them to be added, and (1) is true, or (3) the sequence needs to be tagged as an EST, STS, GSS or HTG.

Entrez Search Field: Keyword [KYWD]
Search Tip: Since keywords are not present in many records, it is best not to search that field. Instead, search All Fields [ALL], the Text Word [WORD] field, or the Title Word [TITL] field, for progressively narrower retrieval.

SOURCE

Free-format information including an abbreviated form of the organism name, sometimes followed by a molecule type. (See section 3.4.10 of the GenBank release notes for more info.)

Entrez Search Field: Organism [ORGN]
Search Tip: For some organisms that have well established common names, such as baker's yeast, mouse, and human, a search for the common name will yield the same results as a search for the scientific name. E.g., a search for "baker's yeast" in the organism field retrieves the same number of documents as "Saccharomyces cerevisiae." This is true because the Organism field is connected to the NCBI Taxonomy Database, which contains cross-references between common names, scientific names, and synonyms for organisms represented in the Sequence databases.

- **Organism**

The formal scientific name for the source organism (genus and species, where appropriate) and its lineage, based on the phylogenetic classification scheme used in the NCBI Taxonomy Database. If the complete lineage of an organism is very long, an abbreviated lineage will be shown in the GenBank record and the complete lineage will be available in the Taxonomy Database. (See also the /db_xref=taxon:nnnn Feature qualifer, below.)

Entrez Search Field: Organism [ORGN]
Search Tip: You can search the Organism field by any node in the taxonomic hierarchy. E.g., you can search for the term "Saccharomyces cerevisiae," "Saccharomycetales," "Ascomycota," etc. to retrieve all the sequences from organisms in a particular taxon.

REFERENCE

Publications by the authors of the sequence that discuss the data reported in the record. References are automatically sorted within the record based on date of publication, showing the oldest references first.

Some sequences have not been reported in papers and show a status of "unpublished" or "in press." When an accession number and/or sequence data has appeared in print, sequence authors should send the complete citation of the article to update@ncbi.nlm.nih.gov and the GenBank staff will revise the record.

Various classes of publication can be present in the References field, including journal article, book chapter, book, thesis/monograph, proceedings chapter, proceedings from a meeting, and patent. The last citation in the References field contains information about the submission itself, rather than a literature citation (see Direct Submission, below).

Entrez Search Field: The various subfields under References are searchable in the Entrez search fields noted below.

- **Authors**

List of authors in the order in which they appear in the cited article.

Entrez Search Field: Author [AUTH]
Search Tip: Enter author names in the form: Lastname AB (without periods after the initials). Initials can be omitted. Truncation can also be used to retrieve all names that begin with a character string, e.g., Richards* or Boguski M*.

- **Title**

Title of the published work, or tentative title of an unpublished work.

Entrez Search Field: Text Word [WORD]
Note: For sequence records, the Title Word [TITL] field of Entrez searches the Definition Line, not the titles of references listed in the record. Therefore, use the Text Word field to search the titles of references (and other text-containing fields).
Search Tip: If a search for a specific term does not retrieve the desired records, try other terms that authors might have used, such synonyms, full spellings, or abbreviations. The 'related records' (or 'neighbors') function of Entrez also allows you to broaden your search by retrieving records with similar sequences, regardless of the descriptive terms used by the submitters.

- **Journal**

MEDLINE abbreviation of the journal name. (Full spellings can be obtained from the PubMed Journal Browser.)

Entrez Search Field: Journal Name [JOUR]
Search Tip: Journal names can be entered as either the full spelling or the MEDLINE abbreviation. You can search the Journal Name field in "Index" mode to see the index for that field, and to select one or more journal names for inclusion in your search.

- **MEDLINE**

MEDLINE unique identifier (UID).

References that include MEDLINE UIDs contain links from the sequence record to the corresponding MEDLINE record. Conversely, MEDLINE records that contain accession number(s) in the SI (secondary source identifier) field contain links back to the sequence record(s).

Entrez Search Field: It is not possible to search the Nucleotide or Protein sequence databases by MEDLINE UID. However, you can search the Literature (PubMed) database of Entrez for the MEDLINE UID, and then link to the associated sequence records.

- **Direct Submission**

Contact information of the submitter, such as institute/department and postal address. This is always the The last citation in the References field. Some older records do not contain the "Direct Submission" reference. However, it is required in all new records.

The Authors subfield contains the submitter name(s), Title contains the words "Direct Submission," and Journal contains the address.

The date in the Journal subfield is the date on which the author prepared the submission. In many cases, it is also the date on which the sequence was received by the GenBank staff, but it is not the date of first public release. If you need to know the latter, send a message to info@ncbi.nlm.nih.gov. We will check the history of the record for you.

Entrez Search Field: Use the Author Field [AUTH] if searching for the author name. Use All Fields [ALL] if searching for an element of the author's address (e.g., Yale University). Note, however, that retrieved records might contain the institution name in a field such as Comment, rather than in the Direct Submission reference, so you might get some false hits.
Search Tip: It is sometimes helpful to search for both the full spelling and an abbreviation, e.g., "Washington University" OR "WashU", since the spelling used by authors might vary.

FEATURES

Information about genes and gene products, as well as regions of biological significance reported in the sequence. These can include regions of the sequence that code for proteins and RNA molecules, as well as a number of other features. (See section 3.4.12 of the GenBank release notes for more info.)

A **complete list of features** is available in the following places:

- Appendix III: Feature keys reference of the DDBJ/EMBL/GenBank Feature Table provides definitions, optional qualifiers, and comments for each feature. An alphabetical list is also available. Appendix IV: Summary of qualifiers for feature keys provides definitions for the Feature qualifiers.
- Sequin Help documentation (scroll down to 'Features' in the table of contents to see an alphabetical list of features with links to descriptions)
- section 3.4.12.1 of the GenBank release notes

The **location of each feature** is provided as well, and can be a single base, a contiguous span of bases, a joining of sequence spans, and other representations. If a feature is located on the complementary strand, the word "complement" will appear before the base span. If the "<" symbol precedes a base span, the sequence is partial on the 5' end (e.g., CDS <1..206). If the ">" symbol follows a base span, the sequence is partial on the 3' end (e.g., CDS 435..915>).

For more information about feature locations, see the Sequin Help Documentation and section 3.4.12.2 of the GenBank release notes.

Entrez Search Field: Feature Key [FKEY]
Search Tip: To scroll through the list of available features, view the Feature Key field in Index mode. You can then select one or more features from the index to include in your query. For example, you can limit your search to records that contain both primer_bind and promoter features.

- **Source**

Mandatory feature in each record that summarizes the length of the sequence, scientific name of the source organism, and Taxon ID number. Can also include other information such as map location, strain, clone,tissue type, etc., if provided by submitter.

Entrez Search Field: All Fields [ALL] can be used to search for some elements in the source field, such as strain, clone, tissue type.

Use the Sequence Length [SLEN] field to search by length, and the Organism [ORGN] field to search by organism name.

Since map location is written as free text and can be represented in a number of ways (e.g., chromosome number, cytogenetic location, marker name, physical map location), it is not directly searchable in the Entrez nucleotides or proteins databases. However, there are a number of resources that allow you to browse and/or search the maps of various genomes.

Taxon

A stable unique identification number for the taxon of the source oganism. A taxonomy ID number is assigned to each taxon (species, genus, family, etc.) in the NCBI Taxonomy Database. See also the Organism field, above.

Entrez Search Field: The Taxonomy ID number is not seachable in the Organism search field of Entrez, but is searchable in the Taxonomy Browser

Note: The /db_xref qualifier is one of many that can be applied to various features. A complete list is available in Appendix IV: Summary of qualifiers for feature keys of the DDBJ/EMBL/GenBank Feature Table, and in section 3.4.12.3 of the GenBank release notes. Appendix III: Feature keys reference shows which qualifiers can be used with specific features (see alphabetical list).

- **CDS**

Coding sequence; region of nucleotides that corresponds with the sequence of amino acids in a protein (location includes start and stop codons). The CDS feature includes an amino acid translation. Authors can specify the nature of the CDS by using the qualifier /evidence=experimental or /evidence=not_experimental.

Submitters are also encouraged to annotate the mRNA feature, which includes the 5'untranslated region (5'UTR), coding sequences (CDS, exon) and 3'untranslated region (3'UTR).

Entrez Search Field: Feature Key [FKEY]
Search Tip: You can use this field to limit your search to records that contain a particular feature, such as CDS. To scroll through the list of available features, view the Feature Key field in Index mode. A complete list of features is also available from the resources noted above.

Protein ID

A protein sequence identification number in the accession.version format that was implemented by GenBank/EMBL/DDBJ in February 1999 (see Version for additional information). Protein IDs consist of three letters followed by five digits, a dot, and a version number. If there is any change to the sequence data (even a single amino acid), the version number will be increased, but the accession portion will remain stable (e.g., AAA98665.1 will change to AAA98665.2).

Entrez Search Field: Can use either the Accession [ACCN] or UID field of the Entrez Proteins database.

GI

"GenInfo Identifier" sequence identification number, in this case, for the protein translation.

The **GI** system of sequence identifiers runs parallel to the **accession.version** system, which was implemented by GenBank, EMBL, and DDBJ in February 1999. Therefore, if the protein sequence changes in any way, it will receive a new GI number, and the suffix of the Protein ID will be incremented by one.

For more information, see the description of Protein ID, above, section 1.3.2 of the GenBank 111.0 release notes, and section 3.4.7 of the current GenBank release notes.

Entrez Search Field: Use the UID field of the **Entrez Proteins** database (the UID field of the Entrez Nucleotides database should be used only for nucleotide sequence identifiers).

Translation

The amino acid translation corresponding to the nucleotide coding sequence (CDS). In many cases, the translations are conceptual. Note that authors can indicate whether the CDS is based on experimental or non-experimental evidence.

Entrez Search Field:It is not possible to search the translation subfield using Entrez. If you want use a string of amino acids as a query to retrieve similar protein sequences, use BLAST instead.

- **Gene** — A region of biological interest identified as a gene and for which a name has been assigned. The base span for the gene feature is dependent on the furthest 5' and 3' features. Additional examples of records that show the relationship between gene features and other features such as mRNA and CDS are AF165912 and AF090832.

  Entrez Search Field: Feature Key [FKEY]
  Search Tip: You can use this field to limit your search to records that contain a particular feature, such as gene. To scroll through the list of available features, view the Feature Key field in Index mode. A complete list of features is also available from the resources noted above.

---

complement — Indicates the feature is located on the complementary strand.

- **Other Features** — Examples of other records that show a variety of biological features; a graphic format is also available for each sequence record, and visually represents the annotated features:

  - **AF165912** (gene, promoter, TATA signal, mRNA, 5'UTR, CDS, 3'UTR) GenBank flat file
  - **AF090832** (protein bind, gene, 5'UTR, mRNA, CDS, 3'UTR) GenBank flat file
  - **L00727** (alternatively spliced mRNAs) GenBank flat file

  A complete list of features is available from the resources noted above.

BASE COUNT — The number of A, C, G, and T bases in a sequence.

ORIGIN

The ORIGIN may be left blank, may appear as `Unreported,' or may give a local pointer to the sequence start, usually involving an experimentally determined restriction cleavage site or the genetic locus (if available). This information is only present in older records.

The sequence data begin on the line immediately below Origin. To view/save the sequence data only, display the record in FASTA format. A description of FASTA format is accessible from the BLAST Web pages.